공학석사학위논문

데이터 기반 센서 보정 및 모바일 센서 배치를 통한 도시 미세먼지 센서네트워크 정확도 향상

Improving Accuracy of Urban Particulate Matter Sensor Network via Data-Driven Sensor Calibration and Mobile Sensor Node Deployment

2020 년 8 월

서울대학교 대학원

기계공학부

이 후 창

Abstract

Improving Accuracy of Urban Particulate Matter Sensor Network via Data-Driven Sensor Calibration and Mobile Sensor Node Deployment

Hoochang Lee
Mechanical Engineering
The Graduate School
Seoul National University

Particulate matter (PM) sensor has been widely deployed to increase spatio-temporal resolution in the urban environment. As a cost-effective PM monitoring solution, low-cost PM sensor ideally stands for dense sensor network nodes. However, low-cost PM sensor remains the doubt of its data reliability. In this paper, we investigate the accuracy of low-cost PM sensor by co-locating a governmental beta attenuation monitor (BAM) for 7.5 months and increase the accuracy with data-driven calibration. We research linear/nonlinear calibration (i.e. multiple linear regression (MLR)/multilayer perceptron (MLP)) and introduce a novel combined calibration. The methods are evaluated by field experiment and are compared with other methods and studies. Also, the data-driven calibration model can utilize for but only a co-located sensor node but also other sensor nodes by using a sensor network. The feasibility of sensor network calibration has been evaluated with experiments.

Keywords: low-cost sensor, particulate matter (PM), sensor network, accuracy, calibration, multilayer perceptron (MLP), multiple linear regression (MLR)
Student Number: *2018-22094*

Contents

List of Figures

Abbreviations

PM	**P**articulate **M**atter
MLR	**M**ultiple **L**inear **R**egression
MLP	**M**ulti **L**ayer **P**erceptron
SMART	**S**elective **M**odel **A**nd **R**esidual **T**reatment
BAM	**B**etta **A**ttenuation **M**onitor
AQI	**A**ir **Q**uality **I**ndex
ReLU	**R**ectified **L**inear **U**nit
PDF	**P**robability **D**ensity **F**unction
MAE	**M**ean **A**bsolute **E**rror
MSE	**M**ean **S**quared **E**rror

Chapter 1

Introduction

Particulate matter (PM) is regarded as the major harmful material which is associated with around 4.2 million death per year worldwide [8, 9]. As it causes various health diseases not only respiratory and cardiovascular disorder but also mental disorders, smart cities have been increasing numerous PM sensor deployment [10]. According to the recent paper, around 10,000 sensor nodes in a city are expected shortly [11]. As the number of nodes is increased for dense PM sensor networks, the low-cost light scattering sensor will be a more essential type of sensor for PM monitoring. However, there are still doubtful points on the data reliability of low-cost sensors. Many researchers have studied the evaluation of low-cost sensors under various global field locations and periods [12–14].

According to the analysis of the researches, low-cost sensors have a limitation on their accuracy and it can be enhanced by the calibration procedures [15–17]. Unfortunately, these researches are mainly based on simple linear regression (two-thirds of total calibrations) and a single calibration model [18]. There is no universal calibration model for all conditions and each calibration model has each advantage and disadvantage. Moreover, the calibration model is affected by geographical location, across different seasons, and the amount of data [19]. In some works, ensemble calibration approaches were applied [20, 21]. However, these researches are restricted on a limited dataset which shared training and test dataset with k-fold cross-validation. In the real-world, one-time calibration is not suit for the nonstationary environment such as seasonality, periodic phenomenon change, and aging effects [22].

In this paper, we introduce a novel ensemble calibration method to cover each model for short periods. Also, we analyze the drift of several periods and build a concept for periodic re-calibration which includes the mobile sensor calibration to adapt drift in each location of static nodes.

Chapter 2

System Description

This section includes the calibration system information on elements and configurations.

2.1 System Elements

2.1.1 Beta Attenuation Monitor: High-End PM Monitoring Reference

The reference system is required to calibrate the low-cost PM sensor and the output from the reference system is regarded as the true value. We selected Beta Attenuation Monitor (BAM) as a reliable reference system since it is the only

continuous measuring regulatory reference that is approved by the Korean Ministry of Environment in Korea. BAM model is PM711 model (Kimoto Inc., Osaka, Japan [1]). The sensing principle of BAM is based on the Bouguer (Lamport-Bear) law that the beta rays exponentially attenuate proportionally to the mass of the solid material. PM is trapped on the filter paper for a while and is penetrated by beta rays. By measuring how the amount of beta rays is absorbed/dissipated, BAM continuously measures the weight concentration of PM.

As shown in Fig. 2.1, PM of a target diameter is classified and is supplied into the measuring part. BAM in the government station is installed in the two systems separately using PM2.5 and PM10 classifier. Humidity, temperature, and flow rate are controlled to stably supply PM measurement samples. These features make BAM high accurate continuous measurement, but it has limitations on slow update rate and expensive costs. These characteristics of BAM has limitations on high spatio-temporal monitoring. There are 40 BAM stations in Seoul and they are operated by the Seoul Research Institute of Public Health. These sparse nodes work as high reliable references for low-cost sensor nodes which has opposite characteristics.

2.1.2 Multi-Sensor Platform: Low-Cost Light Scattering PM Sensor

The light scattering PM sensor is a device that continuously measures the size and number of PM by measuring the scattered light from floating matter. In general,

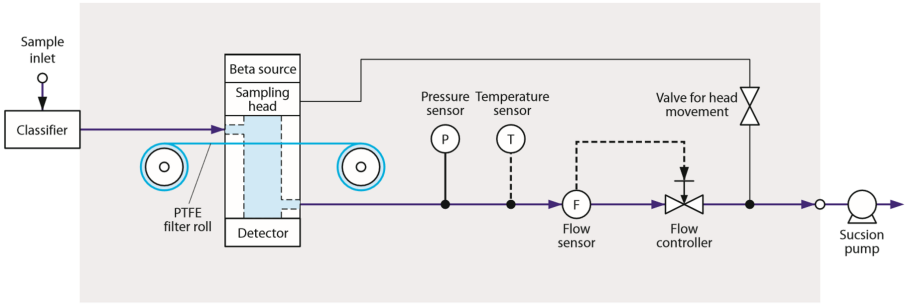


FIGURE 2.1: Functional block diagram of BAM(Figure quoted [1])

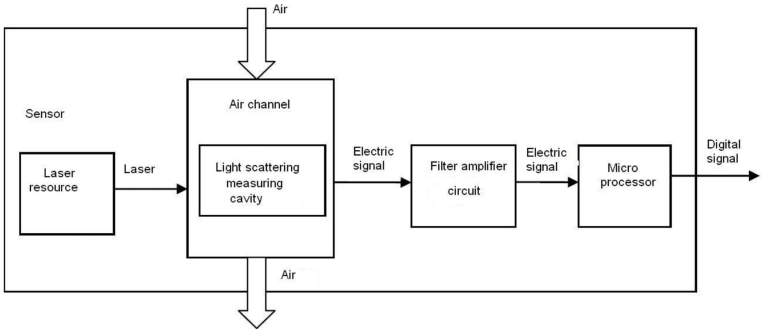


FIGURE 2.2: Functional block diagram of low-cost light scattering PM sensor (Figure quoted [2])

the low-cost sensor is based on light scattering principle and is configured with a small size in the form of a simple measuring device. Not like BAM, it does not include the additional parts for stable measurement, such as a particle seperator, humidity, temperature and flow controller as shown in Fig. 2.2. For these features, it is important for low-cost PM sensor to measure enviroment data and calibrate them according to measured data.

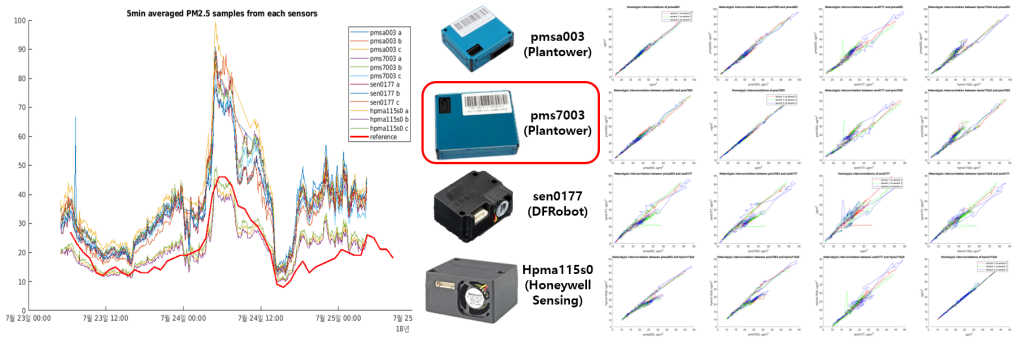


FIGURE 2.3: Prototypes analysis on accuracy (Left) and homo/heterogeneous sample correlation (Right)

Before building a sensor platform for research, we surveyed the other studies to list candidates of PM sensor and built three samples of four kinds prototypes to analyze the homo/heterogeneous sample correlation as shown in Fig. 2.3. Based on the results of prototypes, we selected PMS7003 (Plantower Inc., Beijing, China [2]) and developed a sensor platform for collecting urban PM concentration.

The platform equipped three PM sensors to define sample variation among homogeneous sensors and execute validation of the effectiveness of the sharing calibration model. The sensor platform consists of other environment sensors such as humidity, temperature, light intensity, pressure sensors to evaluate the environmental influence of PM monitoring. Also, the platform includes modules for localization such as inertial measurement unit (IMU) and global positioning system (GPS) to utilize as a mobile sensor node.

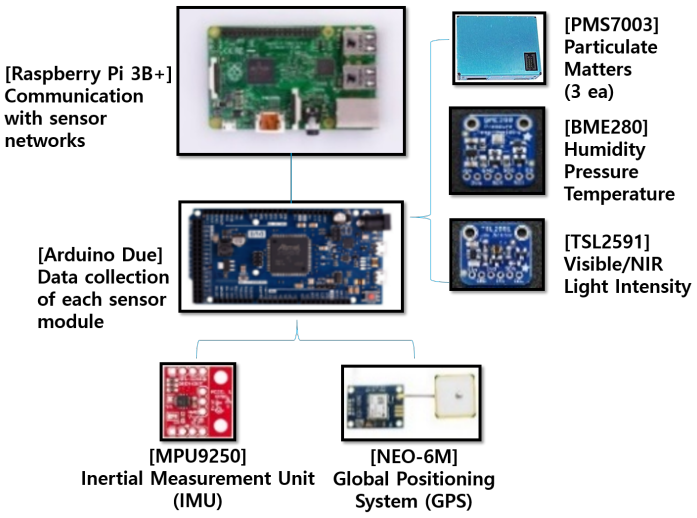


FIGURE 2.4: Module configuration of sensor platform (Figure quoted [3])

As shown in Fig. 2.4, all sensor data collection from sensor modules at a low level is implemented by Arduino Due, and communication with sensor network at a high level is performed by Raspberry Pi 3B+.

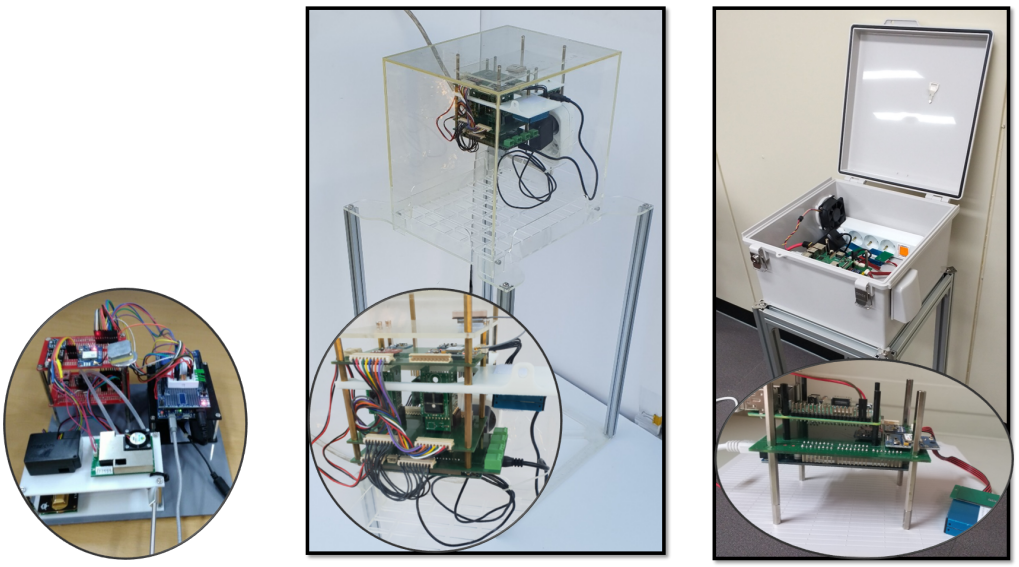


FIGURE 2.5: Multi-sensor platform: prototype (Left), generation 1.0 (Middle), and generation 2.0 (Right)

2.2 System Configuration

2.2.1 Sensor Platform Deployment

The multi-sensor platform was deployed in the government station (6, Sadang-ro 16-gil, Dongjak-gu, Seoul, South Korea) to evaluate the accuracy and build calibration model as shown in Fig. 2.6. nlet of BAM (red circle) and multi-sensor platform (orange circle) are co-located. Data have been collected for 7.5 months (15 Jan. 2019 $\sim$ 4 Sep. 2019).

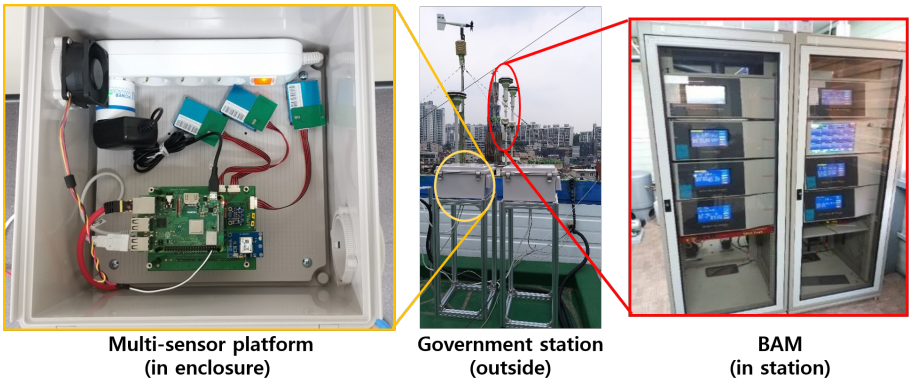


FIGURE 2.6: Sensor platform deployment in government station

2.2.2 Calibration Procedures and Evaluation

The calibration procedure is described in Fig. 2.7. The data from the multi-sensor platform and governmental BAM are followed by pre-condition steps including outlier treatment and time synchronization. In the outlier treatment process, data was rejected if the successive data changes doubled or halved or if any intermittent data were observed from all sensor modules. Since the raw data update rate of the governmental BAM is 5 min, the raw data of the multi-sensor platform were averaged with a 5-min non-overlapping sliding window. Data preprocessing was utilized with Matlab R2018b [23], Python 3, and Pandas (python data manipulation library [24]). The detailed data-driven calibration method will be described in Section 3. The training data and test data are separated into the ratio of 0.8 and 0.2 in sequential order as a default data partitioning condition. The training data use for calibration model build and the

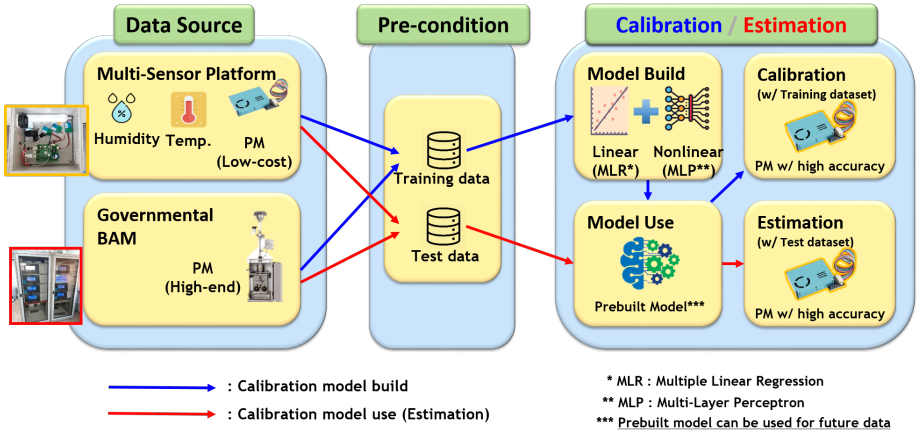


FIGURE 2.7: Overall concepts of calibration procedures (Figure quoted [3])

test data uses for the prebuilt model for estimation. In this paper, calibration means the overall procedure of building a calibration model for estimating the true value, estimation means the procedure of using the prebuilt calibration model without the correction from ground truth data. i.e. BAM data in the test data is not included in the calibration or estimation process, but it is only used in the evaluation of estimation.

The evaluation of estimation is mainly analyzed with four key measures - mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R^2 (coefficient of determination). MAE is a common measure for averaged performance and (R)MSE is a more sensitive measure on the outlier since it proportionally increases weights according to the error magnitude. R^2 measure for how much a variable affects another variable. It's more sensitive to

the total data range. The wider data range tends to be bigger R^2 since the denominator of R^2 is proportionally increased. Mean, standard deviation, quartile, Pearson's correlation coefficient, normalized mean error, slope, intercept are also applied for supplementary descriptive statistics.

TABLE 2.1: Measures for performance analysis

MAE	**MSE**	**RMSE**	$\mathbf{R^2}$
$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i\rvert$	$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$	$1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(\hat{y}_i - \bar{y})^2}$

$y : PM\_reference \qquad \bar{y} : mean\ of\ PM\_reference, \qquad \hat{y} : PM\_estimated$

Chapter 3

Data-Driven Sensor Calibration

In this section, we introduce several data-driven calibration methods and compare the accuracy of those methods.

3.1 Related Studies

3.1.1 w/o Calibration Model

Before the calibration model generation, we analyzed the low-cost PM sensor output without the calibration. It was compared with BAM output for 7.5months.

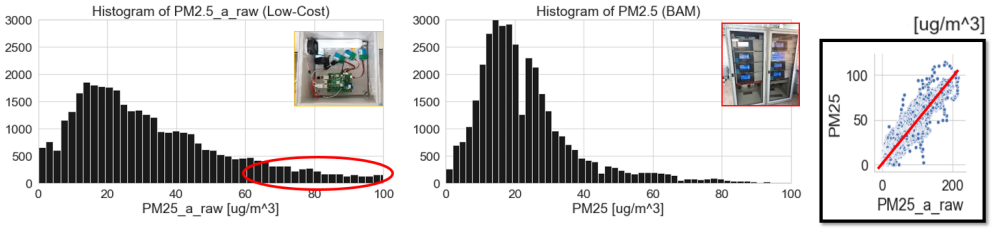


FIGURE 3.1: Histogram for BAM and a low-cost PM sensor

The normalized mean error $[(\bar{y} - \bar{y}_{ref})/\bar{y}_{ref}]$ of low-cost sensor output was 65% of BAM. Also, it had a heavy-tailed distribution and it was over-estimated with almost doubled sensitivity compared to BAM as shown in Fig. 3.1. Three low-cost PM sensors without calibration are also evaluated in scatter and box plot together as shown in Fig. 3.2. Three sample-to-sample variations between homogenous sensors were low and sensors were very strongly correlated with having a correlation coefficient higher than 0.985. These results indicated the low-cost PM sensor had poor accuracy without calibration model and homogenous sensors can share the same calibration model.

3.1.2 Previous Researches

Environmental conditions give an influence on aerodynamic diameter and count of particle. In particular, the influence on humidity and temperature has been continuously studied from the past to the recent studies, and the many calibration models have been developed such as Eq.(3.1) and Eq.(3.2) [25, 26].

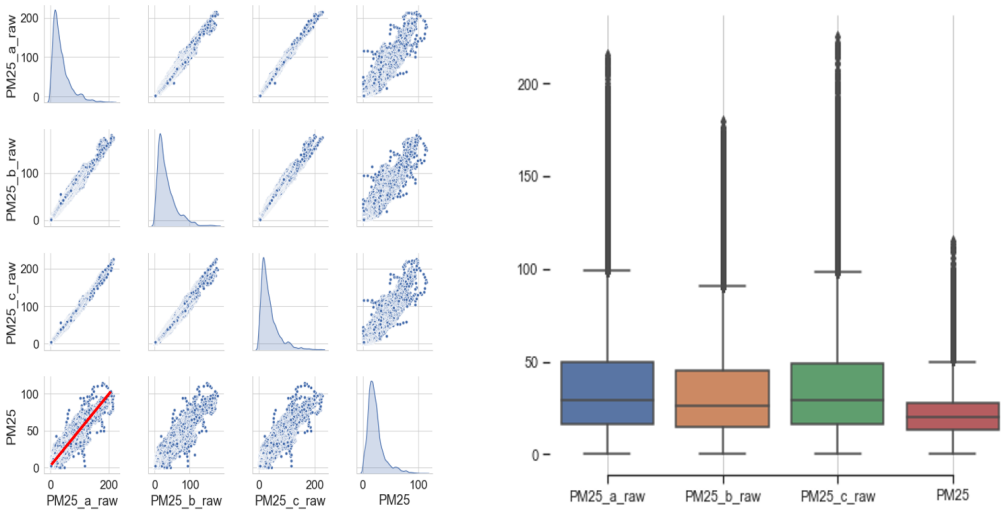


FIGURE 3.2: Scatter plot (Left) and box plot (Right) for BAM and three low-cost PM sensors

$$\hat{y} = \beta_1 + \beta_2 \frac{\rho^2}{1-\rho} y + \beta_0 \tag{3.1}$$

$$\hat{y} = \alpha_1 y + \alpha_2 t + \alpha_0 \tag{3.2}$$

$\hat{y} : PM2.5\_estimated,\ y : PM2.5\_measured,\ \alpha_i, \beta_i : coefficient,\ \rho : RH\_measured,$
$t : Temperature\_measured$

3.2 Linear/Nonlinear Calibration

Low-cost PM sensor has calibration procedure during sensor manufacturing under a controlled environment, but measuring PM concentration is also affected by latent variables such as PM material and its composites. In short, the calibration in manufacturing cannot guarantee the best performance under all use conditions. To increase the accuracy of PM measurement in a specified location, in-field sensor calibration is required. Unfortunately, there is no universal single PM calibration model since each calibration model has each advantage and disadvantage under different geographical locations, different seasons, and the amount of data [19].

Representative linear and nonlinear calibration methods are introduced and their limitations and a new approach to enhance performance are described in followed subsection.

3.2.1 Linear Calibration: Multiple Linear Regression

The major PM calibration method is regarded as linear regression (LR). According to a technical report of the European Joint Research center, two-thirds of PM calibration methods are classified with LR [18].

A low-cost PM sensor is highly affected by environmental conditions. In this paper, multiple linear regression (MLR) uses PM2.5, humidity, and temperature

TABLE 3.1: Chosen coefficients of MLR (80%—training dataset, 20%—test dataset, 5-min. sampling interval condition)

	coef	**std err**	**t**	**P**	**[0.025**	**0.975]**
Intercept	8.4230	0.117	71.986	0.000	8.194	8.652
Raw(a)	0.4490	0.001	426.508	0.000	0.447	0.451
Humidity	-0.0633	0.001	-49.442	0.000	-0.066	-0.061
Temp	0.0255	0.002	10.882	0.000	0.021	0.030

data as explanatory (input) variables and estimates the estimated PM2.5 as a response variable as shown Eq. (3.3). Defining coefficients with the training set is the calibration process of LR as shown Eq. (3.5).

$$\hat{\mathbf{y}} = \mathbf{wx} \tag{3.3}$$

$$\underset{\mathbf{w}}{argmin}(\hat{\mathbf{y}} - \mathbf{y}) \tag{3.4}$$

$\hat{\mathbf{y}} \in \Re^{\mathbf{1}\times\mathbf{r}} : PM2.5\_estimated, \quad \mathbf{y} \in \Re^{\mathbf{1}\times\mathbf{r}} : PM2.5\_ground\ truth,$

$\mathbf{x} \in \Re^{\mathbf{n}\times\mathbf{r}} : input\ variables\_measured, \quad \mathbf{w} \in \Re^{\mathbf{1}\times\mathbf{n}} : coefficient$

$r : \#\ of\ training/test\ instances, \quad n : \#\ of\ input\ variables$

The least-square method was applied for calibration and the coefficients were selected as shown in the table below. Based on the coefficients, the estimation was executed with the test set and Fig. 3.3 shows graphical analysis.

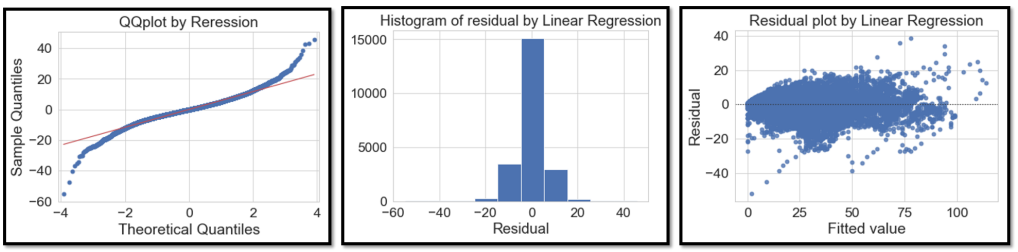


FIGURE 3.3: MLR data plot - QQ plot (Left), Histogram (Middle), and Residual plot (Right)

3.2.2 Nonlinear Calibration: Multilayer Perceptron

These days, machine learning (ML) and artificial neural network (ANN) are broadly researched in many sensor calibration applications. We applied multilayer perceptron (MLP) as a non-linear calibration method since MLP as proven as a universal approximator. MLP consists of an input layer, an output layer, and hidden layers. The calibration model is generated through the process of creating appropriate weights between neurons in adjacent layers as shown in Fig. 3.4. The weights pass through activation function to generate a nonlinear model. In our case, the rectified linear unit (ReLU) activation function was applied as shown (3.5). The learning of weights networks is followed by iterations (epoch x batch) of feedforward and backpropagation steps to minimize cost(error) function.

The hyperparameters of MLP are written as shown in Table. 3.2 and the weight matrix ($\Re^{3\times24}$, $\Re^{24\times24}$, $\Re^{24\times1}$) and bias($3 \times \Re^{1\times24}$) matrix is omitted since it's too big to list in this paper and it does not have any interpretable meaning. Based on the weights, the estimation was executed with the test set and Fig. 3.5

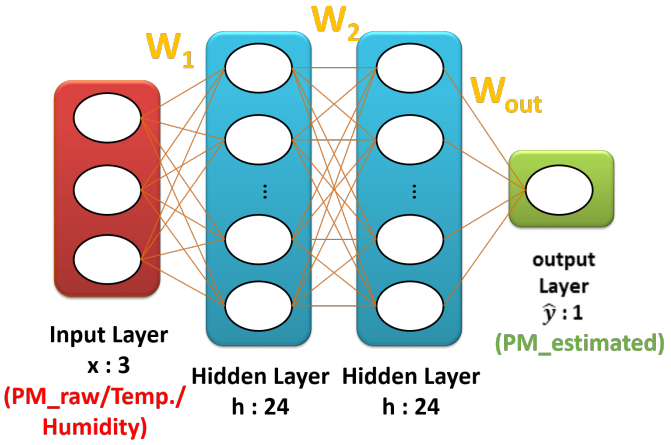


FIGURE 3.4: Architecture of neural networks (fully connected) (Figure quoted [3])

shows graphical analysis and the learning curve in training (including validation data) set is as shown as Fig. 3.6.

$$\hat{\mathbf{y}} = \mathbf{W_{out}} max(0,\ \mathbf{W_2} max(0,\ \mathbf{W_1 x})) \tag{3.5}$$

$\hat{\mathbf{y}} \in \Re^{\mathbf{1\times r}} : PM2.5\_estimated,\ \mathbf{x} \in \Re^{\mathbf{n\times r}} :\ input\ variables\_measured,$
$\mathbf{W_1} \in \Re^{\mathbf{k_a \times n}},\ \mathbf{W_2} \in \Re^{\mathbf{k_b \times k_a}},\ \mathbf{W_{out}} \in \Re^{\mathbf{1\times k_b}} : weights$

$r : \#\ of\ training/test\ instances,\quad n : \#\ of\ input\ variables$

TABLE 3.2: Hyperparameters of MLP (80% - training dataset, 20% - test dataset, 5 min sampling interval condition) (Table quoted [3])

Hidden layer	Neurons / layer	Epoch	Batch	Activation	Dropout rate	Learning rate	Optimizer
2	24	200	32	ReLU	0.2	0.005	Adam

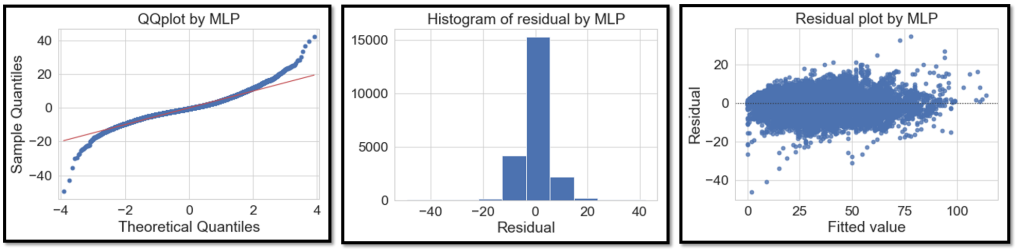


FIGURE 3.5: MLP data plot - QQ plot (Left), Histogram (Middle), and Residual plot (Right)

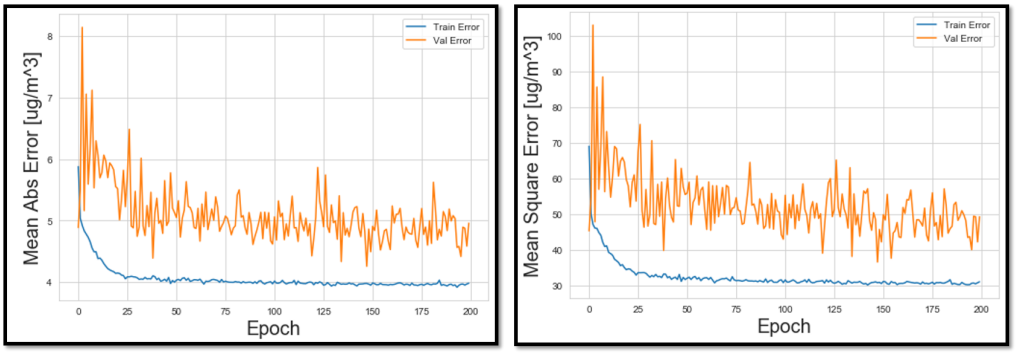
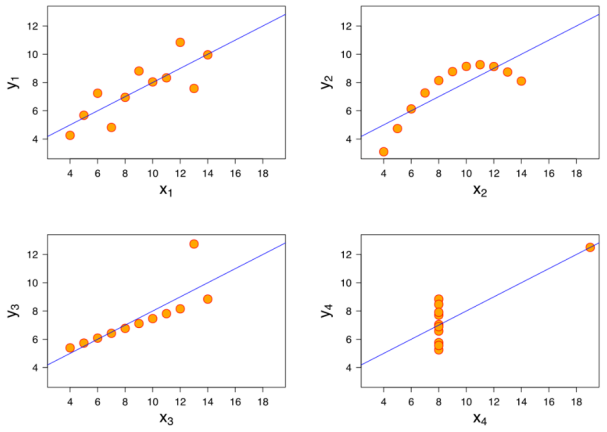


FIGURE 3.6: MLP learning curve in training set

3.2.3 Limitation on Linear/Nonlinear Calibration

LR is the most common technique for finding a best-fit line that represents approximation function and estimates the true values. However, LR is highly



Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

FIGURE 3.7: Anscombe's quartet - dataset I (simple linear) / dataset II (non-linear) / dataset III (linear with outlier) / dataset IV (a high-leverage point). Four datasets have the same mean, variance, Pearson correlation coefficient, R^2, slope, and intercept of the best-fit line (Figure and table quoted [4]).

affected by data's nonlinearity, outliers, and data span. Thes limitations are intuitively interpretable with Anscombe's quartet as shown in Fig. 3.7 [27].

For nonlinear calibration, the approximation model can lower error by increasing model complexity, but this approach is only effective in the training set or seen data. The test set or unseen data with an over-fitted model generates a higher estimation error as shown in Fig. 3.8. So, the generalized model which fits a training set and test set (validation set) together is required or enough training data can avoid the over-fitting issue.

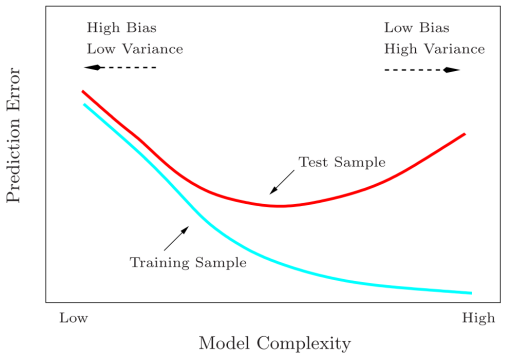


FIGURE 3.8: Prediction Error vs Model complexity in Training/Test sample set (Figure quoted [5]).

3.3 SMART calibration

In the previous section, the benefits and drawbacks of linear/nonlinear calibration models were described. Previous each calibration model has its weak spot in its domain because of the characteristic of linear/non-linear approximation. For instance, LR has a high error at a nonlinear region with best-fit line approximation and MLP has a high error at the over-fitted region. To decrease single deterministic model error, we developed the combined calibration method which has a model selection policy from each segmented domain.

3.3.1 Concepts of Calibration

The overall procedure of segmented model and residual treatment (SMART) calibration is graphically explained as shown in Fig. 3.9. In the model build step,

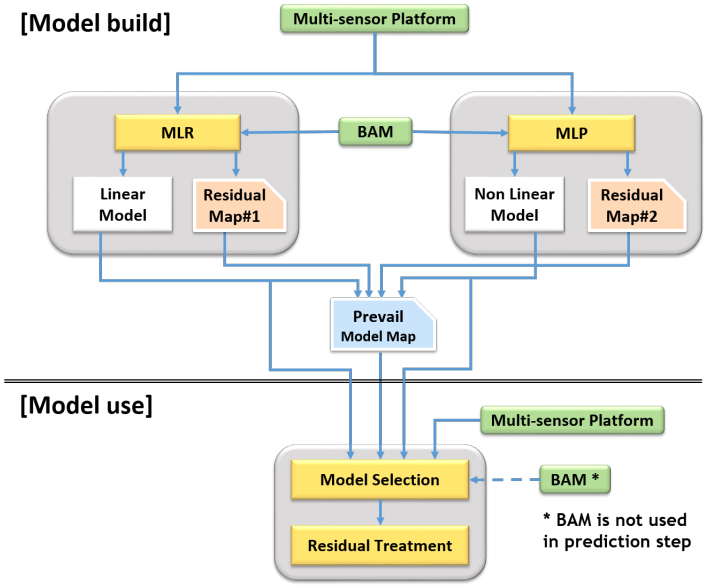


FIGURE 3.9: Overall procedures of SMART Calibration

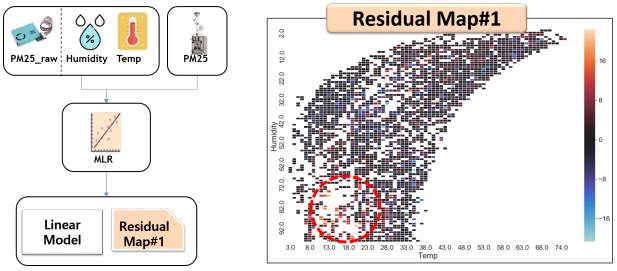


FIGURE 3.10: Residual map generation

each calibration model with its residual map is separately built. The residual map divides the input variable space into separated cells. The separated cell has the local information of corresponding input variable data as shown in Fig. 3.10.

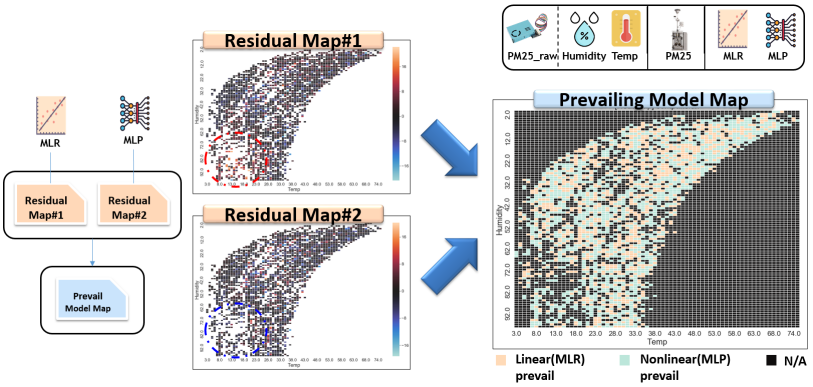


FIGURE 3.11: Prevailing model map generation

The density function of residual in each cell can be assumed as normal distribution since the residual is the error of approximation function. Each cell has its mean and standard deviation of residuals. The residual map contains residual information of cells. After generating each model with its residual map, the prevailing model map is generated by comparing each cell information from residual maps as shown in Fig. 3.11. There are the prevailing calibration model information and residuals' statistics in the cell of the prevailing model map. In the model use step with the test set, the model selection and the residual treatment can be executed with the prevailing model map according to the input variable region as shown in Fig. 3.12.

Above figures are one of examples for SMART calibration, it is not restricted to the number of model/input variable or the type of model/input variable. SMART calibration features the concise procedures and the hierarchical calibration structures. As the performance of SMART calibration is affected by the

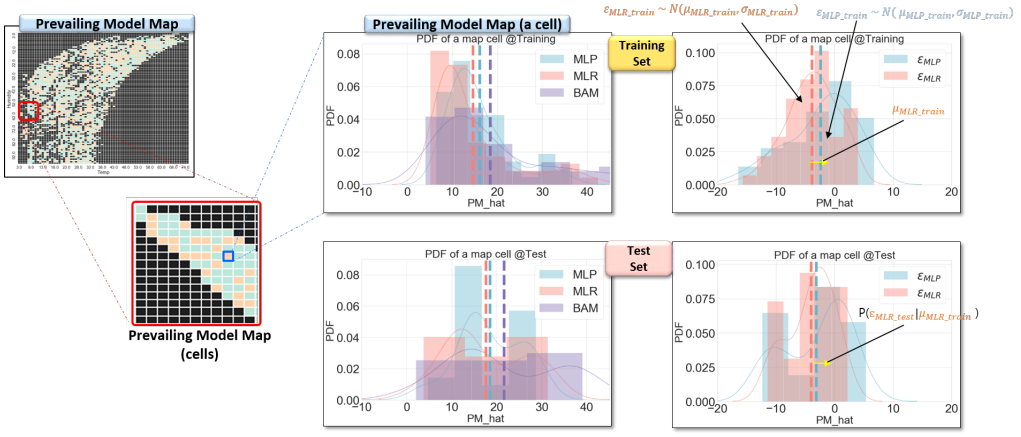


FIGURE 3.12: Prevailing model map (cell information)

number of residuals in each cell, the accuracy can be increased by accumulating more training data. Also, SMART calibration is more effective when high bias and low variance model is applied since residual treatment is more reliable with a consistent estimator.

3.3.2 Procedures of SMART Calibration

This is a SMART calibration example including two calibration models with N input variables and three layers of MLP. It is possible to change the number of models/input variables/hidden layers.

[Model Build Step - Training Dataset]

1. Build a calibration model. ($e.g. MLR : \hat{y} = w_0 + \sum_{i=1}^{N} w_i x_i$)

2. Segment each input space. (i x j matrix)

3. Calculate residuals in each cell (in i x j matrix) according to corresponding data and generate a residual map from the training dataset. (n instances in a cell (i, j cell))

$$\sum_{k=1}^{n} \epsilon_{k[ij]} = \sum_{k=1}^{n} (y_{k[ij]} - \hat{y}_{k[ij]})$$

4. Repeat 1-3 steps for the other calibration models.

5. Compare residual maps for each cell and build a prevailing model map.

prevailing model is selected by $min(\sigma_{\epsilon_{[ij],MLR}}, \sigma_{\epsilon_{[ij],MLP}})$

[Model Use Step - Test Dataset]

6. Estimate truth value from the prevailing model.

$$\tilde{y}_{[ij]} = \hat{y}_{[ij],prevailing\ model}$$

7. Estimate truth value with residual treatment from the prevailing model.

if $\sigma_{\epsilon_{[ij],prevailing\ model}} < \sigma_{\epsilon_{bound}}$

$$\tilde{y}_{[ij]} = \tilde{y}_{[ij]} - \frac{1}{n} \sum_{k=1}^{n} \epsilon_{k[ij],prevailing\ model}$$

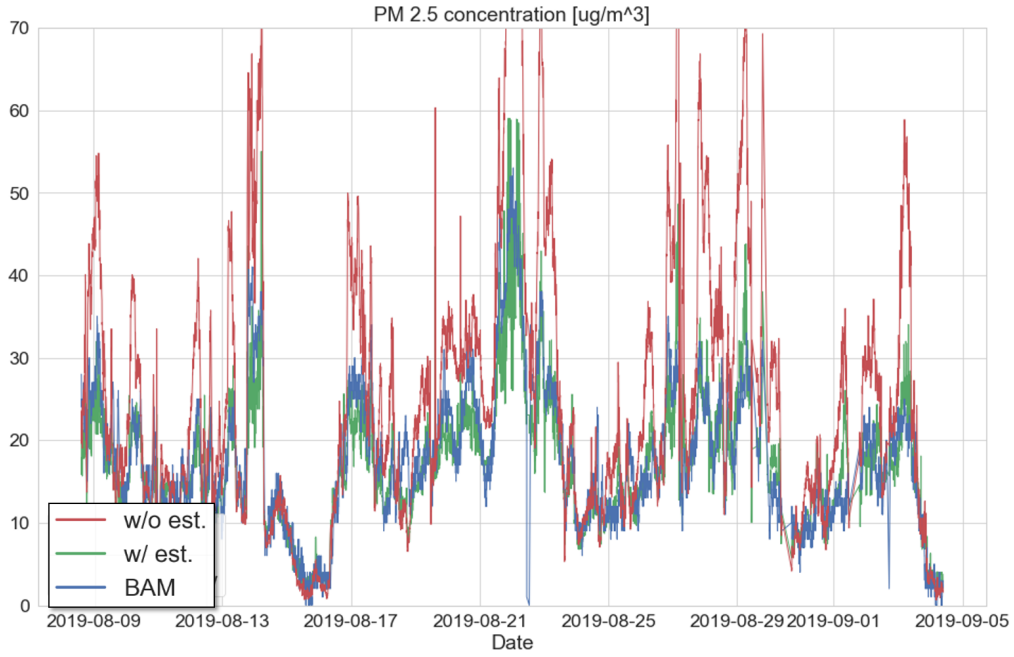


FIGURE 3.13: Data plot for w/ and w/o estimation comparison)

3.4 Experiments and Results

The results of calibration methods were analyzed from various points of view. Firstly, it is compared with the raw data (w/o estimation) and BAM data in the given preprocess condition (80 % training and 20 % test). SMART calibration result is plotted with BAM output and raw output as shown in Fig. 3.13. Output w/ estimation decreased MAE from 9.06 to 2.79 $[ug/m^3]$. Secondly, several calibration results are evaluated together with a scatter plot and box plot as shown in Fig. 3.14.

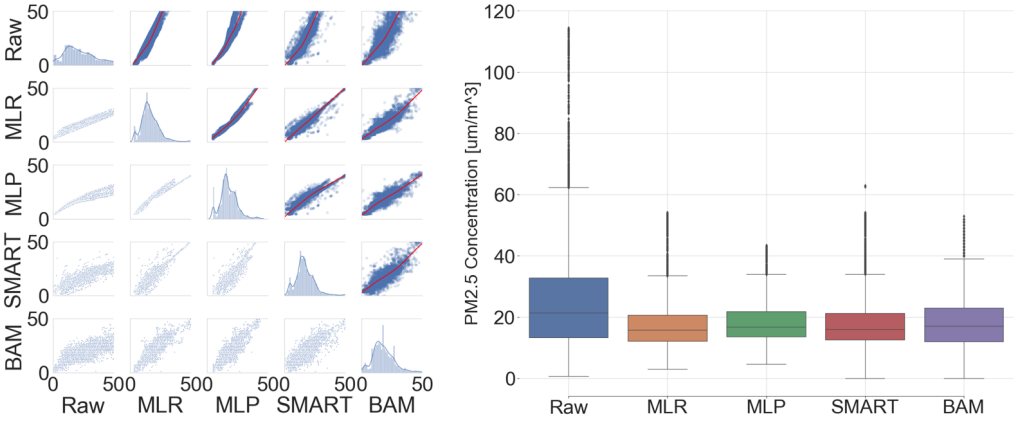


FIGURE 3.14: scatter plot (Left) and box plot (Right) for w/ and w/o estimation comparison (Figure quoted [3])

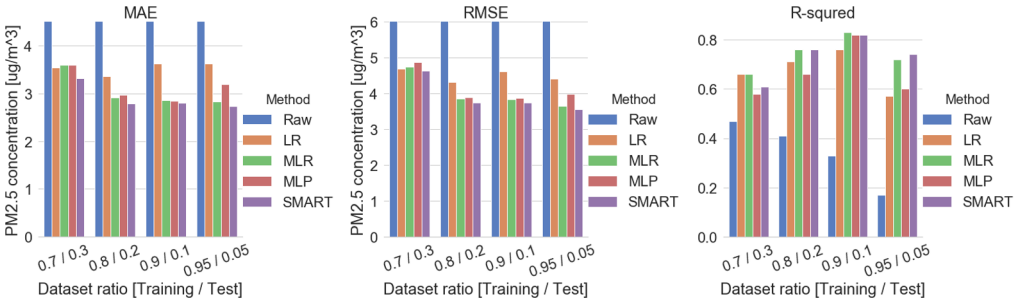


FIGURE 3.15: Key measures for w/ and w/o estimation comparison (Figure quoted [3])

The calibration results are compared with the key measures in various data partitioning ratios as shown in Fig. 3.15. The Numerical analysis is written in the Table. 3.3 with key measures of other calibration models.

TABLE 3.3: Key measures analysis of various calibration methods. polynomial linear regression (PLR), Lasso/Ridge regression (Lasso/Ridge), decision tree(DT), random forests (RF), gradient boosting (GB), extreme GB (XGB), and light GB (LGB) (Figure quoted [3])

Data set Ratio	Measures	Raw	LR	MLR	MLP	SMART	PLR	Lasso	Ridge	DT	RF	GB	XGB	LGB
70%/ 30%	**MAE**	8.92	3.54	3.60	3.60	3.32	3.31	3.40	3.60	4.00	3.00	3.00	**2.98**	3.15
	MSE	182.31	21.99	22.49	23.70	21.56	19.21	20.71	22.49	30.18	16.69	16.45	**16.26**	17.82
	$\mathbf{R}^2$	0.47	0.66	0.66	0.58	0.61	0.65	0.68	0.66	0.69	0.78	0.77	0.77	0.74
80%/ 20%	**MAE**	9.06	3.36	2.91	2.97	**2.79**	2.94	2.92	2.91	3.39	2.85	2.88	2.79	2.84
	MSE	196.35	18.70	14.84	15.20	14.02	14.80	14.98	14.84	21.24	14.43	14.58	**13.80**	14.26
	$\mathbf{R}^2$	0.41	0.71	0.76	0.66	0.76	0.75	0.75	0.76	0.71	0.77	0.78	**0.78**	0.79
90%/ 10%	**MAE**	11.67	3.62	2.86	2.84	**2.80**	2.87	2.85	2.86	3.81	2.95	2.85	2.85	2.95
	MSE	311.90	21.31	14.73	15.06	**14.05**	14.67	14.61	14.73	26.74	15.11	14.78	14.71	15.54
	$\mathbf{R}^2$	0.33	0.76	0.83	0.82	0.82	0.83	0.83	0.83	0.72	0.81	0.83	**0.84**	0.83
95%/ 5%	**MAE**	10.07	3.63	2.83	3.19	**2.74**	2.81	2.80	2.83	3.33	2.86	2.84	2.86	2.89
	MSE	194.54	19.44	13.34	15.92	**12.75**	13.21	13.01	13.34	19.20	13.37	13.88	14.07	14.14
	$\mathbf{R}^2$	0.17	0.57	0.72	0.60	0.74	0.72	0.71	0.72	0.67	0.71	**0.75**	0.74	0.74

3.4.1 Comparison w/ Other Calibration Methods

The SMART calibration result was compared with other regression-based models and classification and regression trees (CART) models including the latest ensemble learning methods such as XGB and LGB. The hyperparameters of each model have been checked exhaustively over a pre-defined set of hyperparameters with a cross-validated grid search algorithm as shown in Table. 3.3 and Fig. 3.16.

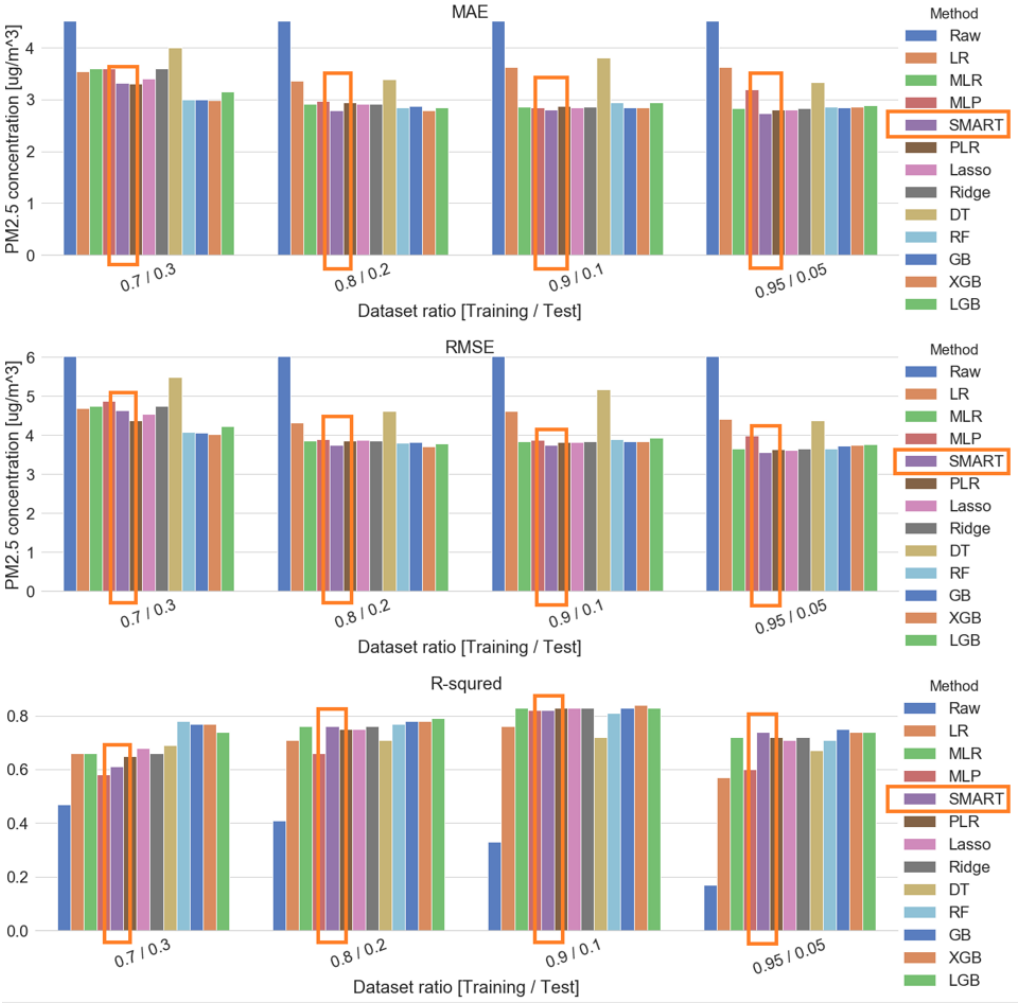


FIGURE 3.16: Key measures analysis of various calibration methods (GridsearchCV (10)). PLR (degree:2) / Lasso (alpha:5) / Ridge (alpha:100) / DT (max depth = 12, min samples split =16) / RF (max depth = 6, min samples leaf = 8, min samples split = 24, n estimators = 500) / GB (learning rate = 0.05, n estimators = 200) / XGB (colsample bytree = 1, learning rate = 0.05, n estimators = 200, subsample = 0.3) / LGB (colsample bytree = 0.5, learning rate = 0.05, n estimators = 500, num leaves = 4, reg lambda = 10, subsample = 0.3) (Figure quoted [3])

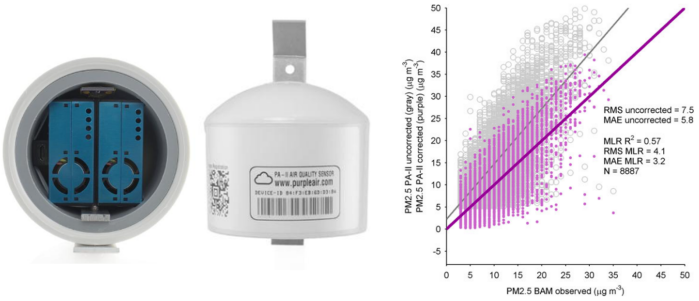


FIGURE 3.17: PA-II (Left) (Figure quoted [6]) and long-term evaluation and calibration (Figure quoted [7])

3.4.2 Comparison w/ Other Studies

Magi et.al. 2019 had long-term in-field calibration research for 16 months in North Carolina, USA [7]. They conducted a long-term experiment with a commercial product, PA-II (Purple Air Inc., UT, USA [6]), and evaluated 1-hour sampling interval basis with 90% training dataset and 10% test dataset (10,000 Monte Carlo). PM 2.5, humidity, and temperature were measured as input variables and BAM 1020 (Met One Inc., OR, USA [28]) was used as ground truth. Our dataset was applied to 1-hour sampling interval and sequantial 90% training dataset with SMART calibration. The results were compared as shown in Table 3.4. Our group's result with sequential SMART calibration showed better accuracy having a sharp decline in errors.

TABLE 3.4: Performance comparison by group

Category	Measures	Other Group [MLR]	Our Group [SMART]
w/ est. (change)	**MAE [ug/m3]**	3.2 (-2.6)	2.8 (-8.6)
	RMSE[ug/m3]	4.1 (-3.4)	3.7 (-13.6)
	R^2	0.57 (+0.03)	0.81 (+0.47)
w/o est.	**MAE [ug/m3]**	5.8	11.4
	RMSE[ug/m3]	7.5	17.3
	R^2	0.54	0.34

3.4.3 Further Analysis of Calibration Model

PM concentration is known to have seasonal patterns and it requires sufficient long-term data to avoid over-fitting issues. According to the latest analysis, a test dataset of at least a year is needed to capture all seasonal phenomenons [19]. Also, the light-scattering PM sensor is vulnerable to be mechanically contaminated with particle on the lens. These limitations cause concept drift from the calibration model and it can be resolved with periodic on/offline re-calibration [29]. To check the approximate level of concept drift, we compared three differently separated test set from our 7.5 months dataset (1 month/3 months/5 months data as test set). The mean of MAE was increased as time passes and the minimum of MAE was shown in right after calibration model applied. Also, MAE was largely observed under the unseen seasonal influence.

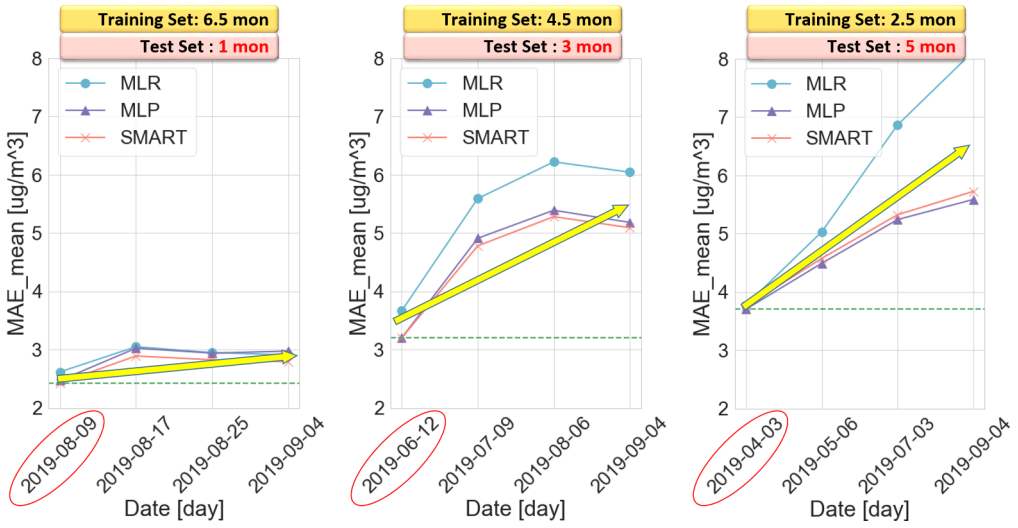


FIGURE 3.18: Concept drift analysis

Chapter 4

Sensor Network Calibration

In this section, the phased calibration method concept, Hybrid Calibration, is introduced to increase the accuracy of sensor networks. The hybrid calibration is defined as the combination of an online calibration and an offline calibration. The online calibration performs the data-driven calibration model translation to other nodes through sensor network to increase accuracy. However, this calibration method has a limitation on including the local variation. To overcome this limitation, the offline calibration uses a mobile sensor node by locally approaching nodes to minimize the local variation of the transferred calibration model. In this paper, the online calibration methods are called transfer calibration, and the offline calibration is called rendezvous calibration.

4.1 Related study

4.1.1 Sensor Network Calibration

There are various researches on sensor network calibration. One of the simple ways is sharing the ground truth measurement to re-calibrate sensor nodes when some ratio of total nodes is below a certain concentration level [30]. However, this method is assumed the nodes are exposed under the same phenomenon w/o local pollutants. It can minimize sensor drift, but it may cause miscalibration due to strong assumptions. Beyond directly sharing the ground truth measurement, there is another approach of sharing the calibration model which is generated by the in-field test. It's mainly used in chemical gas sensors since a sensor has a cross-sensitivity on other gases and suffers limitation on the test of gas composite in their manufacturing process. Cheng et al. applied transfer calibration to real-world PM sensor networks [31]. However, the accuracy improvement was not clearly described in the paper. e.g. the estimated output was described as a not continuous variable but 8 levels categorical index (i.e. level 1: 0-35, level 2: 35-75, level 3: 75-115 $[ug/m^3]$) and compared with ground truth according to the analysis of confusion matrix.

Also, there is another calibration approach of sensor network calibration by using a mobile sensor node. The mobile sensor node physically approaches to compare and share the information of nodes. The previous calibration study by the mobile sensor was a single-hop algorithm via adjacent nodes [32]. Also, a multi-hop

algorithm was developed by preventing the issue of diluting the calibration parameter [17] but both researches are about chemical sensors, not for PM sensors.

4.1.2 Mobile Sensor Node

A mobile sensor node offers the opportunity of PM measuring since it can travel or change deployment location. It creates increased spatial resolution in transit and on a new location. Also, it provides comparatively duplicated information when co-locating with static nodes for maintenance or calibration.

4.2 Transfer Calibration

4.2.1 Concepts of Transfer Calibration

Transfer calibration shares the model to other nodes in sensor network to increase the accuracy as shown in Fig. 4.1. The major benefit of transfer calibration is sharing and using a calibration model for nodes which does not have a reference node. Using the calibration model of a master static node (S_1) can be valuable under the condition of the same characteristics of sensors. The low-cost sensors were very strongly correlated and were analyzed with an extremely high correlation coefficient ranging from 0.985 to 0.997 in the same location [3]. Based on this analysis, the built model from data-driven calibration can be shared with other sensor nodes through sensor network.

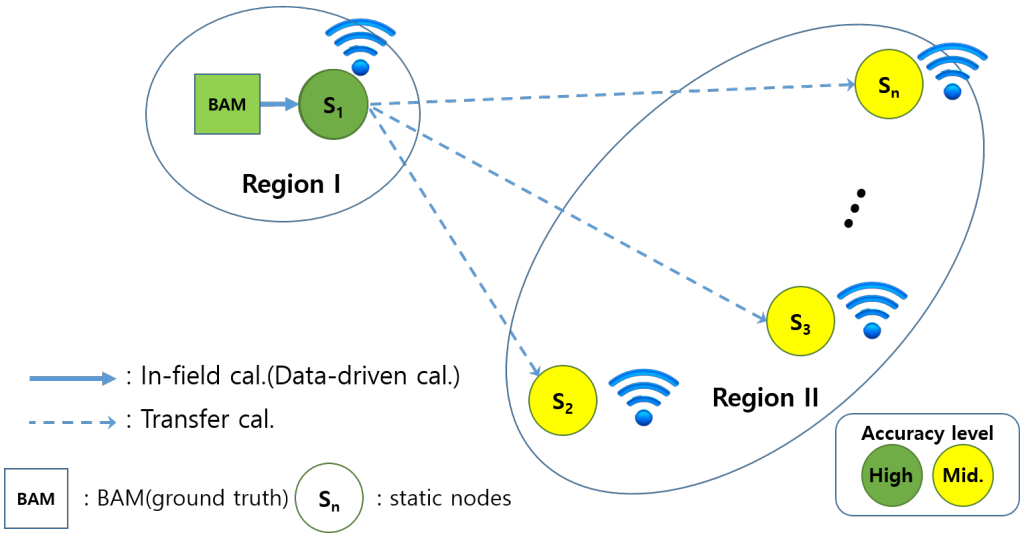


FIGURE 4.1: Concept diagram: transfer calibration

As the applied calibration model has the drift of calibration, the regularly transferring calibration model is required. The data-driven calibration model can be continuously updated by co-locating sensor platform with BAM.

4.3 Rendezvous Calibration

Basically, the online calibration increases accuracy of sensor networks. However, the online calibration method may have a limitation on covering spatial differences of PM composites which is one of the latent variables of PM concentration. So, offline calibration is required to support this limitation of online calibration. The mobile sensor node is used for offline calibration as shown in Fig. 4.2. The

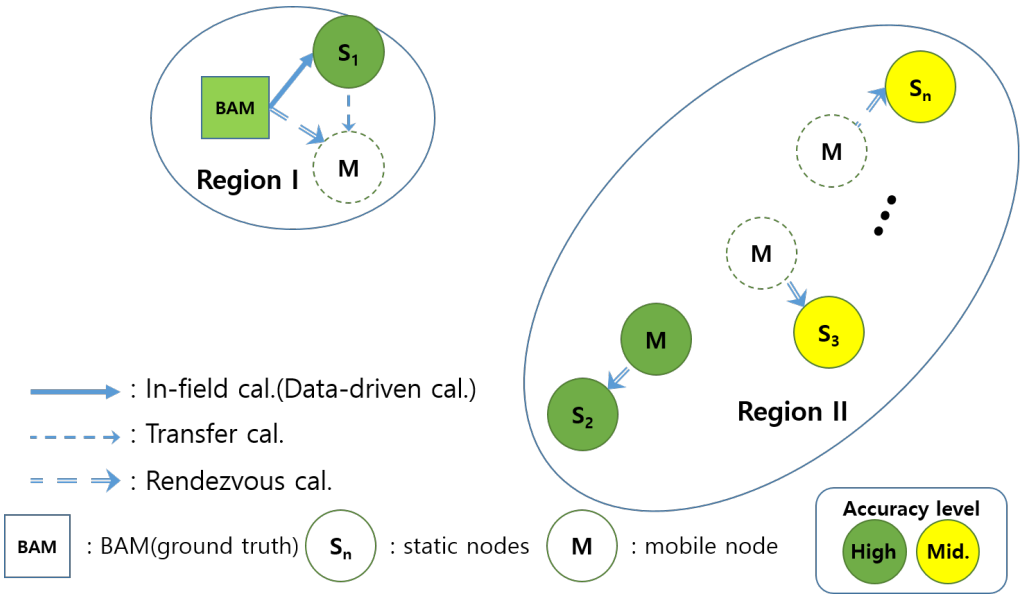


FIGURE 4.2: Concept diagram: rendezvous calibration

mobile node also applied the transferred calibration model and the model updated by the observation of BAM when it is located with BAM. On the contrary, the mobile node can be regarded as a reference node when it has a successful rendezvous in other static nodes. For this offline correction, Kalman filter model-based approach is under consideration.

4.4 Experiments and Results

Two multi-sensor platforms were deployed in two government stations (around 5km straight distance: 6, Sadang-ro 16-gil, Dongjak-gu, Seoul, South Korea, and

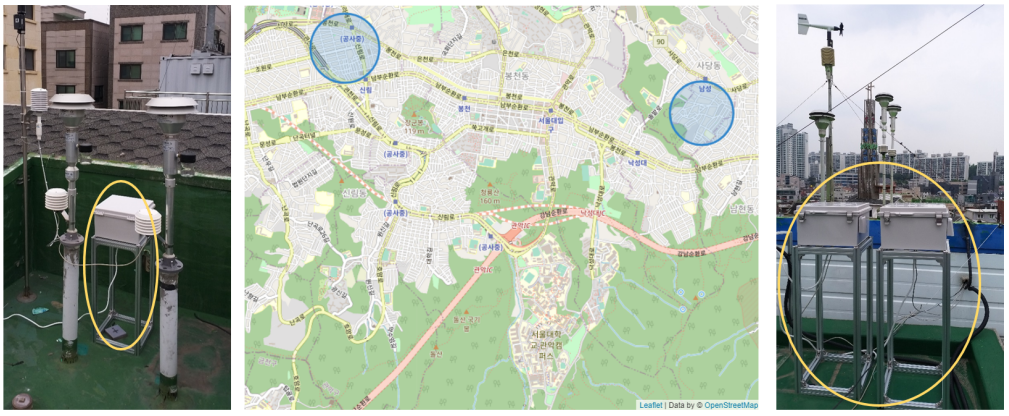


FIGURE 4.3: Sensor platform deployment in two govenment station. Gwanak-gu(Left) and Dongjak-gu(Right)

14, Sillimdong-gil, Gwanak-gu, Seoul, South Korea) to check the feasibility of sensor network calibration as shown in Fig. 4.3. Data have been collected in each static node for 1.5 months (24 May. 2020 $\sim$ 8 Jul. 2020) The calibration model build was executed in Dongjak-gu on 21st June, 28th June, and 5th July. The models were transferred to a static node in Gwanak-gu.

As shown in Table. 4.1, MAE was decreased from 19.82 to 7.01(SMART)$[ug/m^3]$. Also when the number of transfer calibration was increased, MAE was decreased for every estimation models.

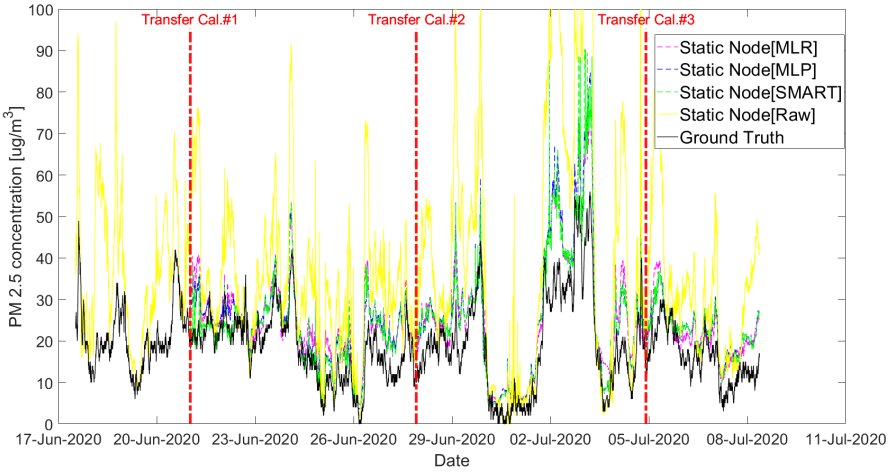


FIGURE 4.4: Transfer calibration result

TABLE 4.1: MAE comparison $[ug/m^3]$

Cases	**w/o est.**	**w/ est.**		
	Raw	**MLR**	**MLP**	**SMART**
Transfer Cal.#1	19.8274	7.4631	7.3139	7.0135
Transfer Cal.#1 & #2	19.8274	6.9912	7.1600	6.9080
Transfer Cal.#1 & #2 & #3	19.8274	6.8067	6.9414	6.7069

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this paper, we analyze the accuracy of the low-cost PM sensor concentration in the urban environment for 7.5 months. The raw data (w/o estimation) had 65% of normalized mean error and the three PM sensor showed an extremely high correlation from 0.985 to 0.997. It indicates the low-cost sensor requires the calibration process and the common calibration can be shared with other nodes in sensor network. We applied the linear/nonlinear calibration (i.e. MLR and MLP) and introduced SMART calibration which selects model and compensates residuals according to the input data region. By applying SMART calibration, MAE was decreased from 9.06 to 2.79 $[ug/m^3]$. SMART calibration compared with

other regression-based models and classification and regression trees (CART) models including the latest ensemble learning methods such as XGB and LGB. As the data sample size increased, the accuracy of SMART calibration increased. We found the concept drift from the built estimation model, it can be solved by the re-calibration from sensor networks. Also we introduce the concept for sensor network calibration and check the feasibility of concepts.

5.2 Future Work

Based on our research, a calibration model can be continuously updated and improved by co-locating a single multi-sensor platform with BAM and can be transferred toward all nodes in a sensor network to calibrate the entire nodes. This approach is the base concept of an online calibration for low-cost sensors. Also, mobile calibration method can be used for spatial variation handling. This successive calibration will feature both an entire online calibration and an individual offline calibration to increase the accuracy of the urban PM sensor network.

Bibliography

[1] Kimoto Inc. Website for detail information(datasheet and technical documents. Available online : https://www.kimoto-electric.co.jp/english/product/air/700.html#lineup. (accessed on 3 January 2020).

[2] Plantower Inc. Website for detail information(datasheet and technical documents. Available online : http://www.plantower.com/en/list/?118_1.html. (accessed on 3 January 2020).

[3] Hoochang Lee, Jiseock Kang, Sungjung Kim, Yunseok Im, Seungsung Yoo, and Dongjun Lee. Long-term evaluation and calibration of low-cost particulate matter (PM) sensor. *Sensors (Switzerland)*, 20(13):1–23, 2020. ISSN 14248220. doi: 10.3390/s20133617.

[4] Wikipedia - Anscombe's quartet. Available online : https://en.wikipedia.org/wiki/Anscombe%27s_quartet. (accessed on 18 December 2019).

[5] Klaus Nordhausen. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition by Trevor Hastie, Robert Tibshirani, Jerome Friedman. *Int. Stat. Rev.*, 77(3):482, 2009. ISSN 03067734. doi: 10.1111/j.1751-5823.2009.00095_18.x.

[6] Purple Air Inc. Website for detail information(datasheet and technical documents. Available online : https://www2.purpleair.com/products/purpleair-pa-ii. (accessed on 3 January 2020).

[7] Brian I. Magi, Calvin Cupini, Jeff Francis, Megan Green, and Cindy Hauser. Evaluation of PM2.5 measured in an urban setting using a low-cost optical particle counter and a Federal Equivalent Method Beta Attenuation Monitor. *Aerosol Sci. Technol.*, 0(0):1–13, 2019. ISSN 15217388. doi: 10.1080/02786826.2019.1619915. URL https://doi.org/10.1080/02786826.2019.1619915.

[8] Richard Burnett, Hong Chen, and Szyszkowicz. Global estimates of mortality associated with longterm exposure to outdoor fine particulate matter. *Proc. Natl. Acad. Sci. U. S. A.*, 115(38):9592–9597, 2018. ISSN 10916490. doi: 10.1073/pnas.1803222115.

[9] World Health Organization (WHO). *Evolution of WHO air quality guidelines*. 2017. ISBN 9789289052306.

[10] Suji Lee, Whanhee Lee, Dahye Kim, Ejin Kim, Woojae Myung, Sun-young Kim, and Ho Kim. Short-term PM 2 . 5 exposure and emergency hospital

admissions for mental disease. 171(March 2018):313–320, 2019. doi: 10.1016/j.envres.2019.01.036.

[11] Naser Hossein Motlagh, Tuukka Petaja, Markku Kulmala, Sasu Trachoma, Eemil Lagerspetz, Petteri Nurmi, Xin Li, Samu Varjonen, Julien Mineraud, Matti Siekkinen, Andrew Rebeiro-Hargrave, and Tareq Hussein. Toward Massive Scale Air Quality Monitoring. *IEEE Commun. Mag.*, 58(2):54–59, 2020. ISSN 0163-6804. doi: 10.1109/mcom.001.1900515.

[12] Florentin M.J. Bulot, Steven J. Johnston, Philip J. Basford, Natasha H.C. Easton, Mihaela Apetroaie-Cristea, Gavin L. Foster, Andrew K.R. Morris, Simon J. Cox, and Matthew Loxham. Long-term field comparison of multiple low-cost particulate matter sensors in an outdoor urban environment. *Sci. Rep.*, 9(1):1–13, 2019. ISSN 20452322. doi: 10.1038/s41598-019-43716-3. URL http://dx.doi.org/10.1038/s41598-019-43716-3.

[13] Anondo Mukherjee, Levi G. Stanton, Ashley R. Graham, and Paul T. Roberts. Assessing the utility of low-cost particulate matter sensors over a 12-week period in the Cuyama valley of California. *Sensors (Switzerland)*, 17(8), 2017. ISSN 14248220. doi: 10.3390/s17081805.

[14] Hai Ying Liu, Philipp Schneider, Rolf Haugen, and Matthias Vogt. Performance assessment of a low-cost PM 2.5 sensor for a near four-month period in Oslo, Norway. *Atmosphere (Basel).*, 10(2), 2019. ISSN 20734433. doi: 10.3390/atmos10020041.

[15] Yi Gao, Wei Dong, Kai Guo, Xue Liu, Yuan Chen, Xiaojin Liu, Jiajun Bu, and Chun Chen. Mosaic: A low-cost mobile sensing system for urban air quality monitoring. In *Proc. - IEEE INFOCOM*, volume 2016-July. Institute of Electrical and Electronics Engineers Inc., jul 2016. ISBN 9781467399531. doi: 10.1109/INFOCOM.2016.7524478.

[16] BALZ MAAG. *(Thesis) Air Quality Sensor Calibration and its Peculiarities*. PhD thesis, 2019. URL https://doi.org/10.3929/ethz-a-010025751.

[17] Balz Maag, Zimu Zhou, Olga Saukh, and Lothar Thiele. SCAN: Multi-Hop Calibration for Mobile Sensor Arrays. *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, 1(2):1–21, 2017. ISSN 2474-9567. doi: 10.1145/3090084.

[18] F. Karagulian, M. Gerboles, and Basr. *Review of sensors for air quality monitoring*. 2019. ISBN 9789276092551. doi: 10.2760/568261. URL https://publications.jrc.ec.europa.eu/repository/bitstream/JRC116534/kjna29826enn.pdf.

[19] Francesco Concas, Julien Mineraud, Eemil Lagerspetz, Samu Varjonen, Kai Puolamäki, Petteri Nurmi, and Sasu Tarkoma. Low-Cost Outdoor Air Quality Monitoring and In-Field Sensor Calibration. (February), 2019. URL http://arxiv.org/abs/1912.06384.

[20] Yuxiang Lin, Wei Dong, and Yuan Chen. Calibrating Low-Cost Sensors by a Two-Phase Learning Approach for Urban Air Quality Measurement. *Proc.*

ACM Interactive, Mobile, Wearable Ubiquitous Technol., 2(1):1–18, 2018. ISSN 2474-9567. doi: 10.1145/3191750.

[21] José María Cordero, Rafael Borge, and Adolfo Narros. Using statistical methods to carry out in field calibrations of low cost air quality sensors. *Sensors Actuators, B Chem.*, 267(2):245–254, 2018. ISSN 09254005. doi: 10.1016/j.snb.2018.04.021. URL https://doi.org/10.1016/j.snb.2018.04.021.

[22] Gregory Ditzler, Manuel Roveri, Cesare Alippi, and Robi Polikar. Learning in Nonstationary Environments: A Survey. *IEEE Comput. Intell. Mag.*, 10 (4):12–25, 2015. ISSN 1556603X. doi: 10.1109/MCI.2015.2471196.

[23] Matlab R2018b. Available online : https://www.mathworks.com/. (accessed on 3 January 2020).

[24] Pandas. Available online : https://pandas.pydata.org/. (accessed on 3 January 2020).

[25] Leigh R. Crilley, Marvin Shaw, Ryan Pound, Louisa J. Kramer, Robin Price, Stuart Young, Alastair C. Lewis, and Francis D. Pope. Evaluation of a low-cost optical particle counter (Alphasense OPC-N2) for ambient air monitoring. *Atmos. Meas. Tech.*, 11(2):709–720, 2018. ISSN 18678548. doi: 10.5194/amt-11-709-2018.

[26] Tongshu Zheng, Michael H Bergin, Karoline K Johnson, Sachchida N Tripathi, Shilpa Shirodkar, Matthew S Landis, Ronak Sutaria, and David E

Carlson. Field evaluation of low-cost particulate matter sensors in high- and low-concentration environments. *Atmos. Meas. Tech.*, 11(8):4823–4846, 2018. ISSN 18678548. doi: 10.5194/amt-11-4823-2018.

[27] W.J. Dixon and A.M. Mood. Graphs in Statistical Analysis Author (s): F . J . Anscombe Source : The American Statistician , Vol . 27 , No . 1 (Feb ., 1973), pp . 17-21 Published by : Taylor & Francis , Ltd . on behalf of the American Statistical Association Stable URL : https:/. 27(1):17–21, 1973.

[28] Metone Inc. Website for detail information(datasheet and technical documents. Available online : https://metone.com/products/bam-1020/. (accessed on 3 January 2020).

[29] Jose M. Barcelo-Ordinas, Messaoud Doudou, Jorge Garcia-Vidal, and Nadjib Badache. Self-calibration methods for uncontrolled environments in sensor networks: A reference survey. *Ad Hoc Networks*, 88:142–159, 2019. ISSN 15708705. doi: 10.1016/j.adhoc.2019.01.008. URL https://doi.org/10.1016/j.adhoc.2019.01.008.

[30] Theofylaktos Pieri and Michalis P. Michaelides. Air pollution monitoring in lemesos using a wireless sensor network. *Proc. 18th Mediterr. Electrotech. Conf. Intell. Effic. Technol. Serv. Citizen, MELECON 2016*, (April):18–20, 2016. doi: 10.1109/MELCON.2016.7495468.

[31] Yun Cheng, Xiucheng Li, Zhijun Li, Shouxu Jiang, Yilong Li, Ji Jia, and Xiaofan Jiang. AirCloud: a cloud-based air-quality monitoring system for

everyone [GP]. *Proc. 12th ACM Conf. Embed. Netw. Sens. Syst. - SenSys '14*, pages 251–265, 2014. doi: 10.1145/2668332.2668346. URL http://dl.acm.org/citation.cfm?doid=2668332.2668346.

[32] Adrian Arfire, Ali Marjovi, and Alcherio Martinoli. Model-based rendezvous calibration of mobile sensor networks for monitoring air quality. *2015 IEEE SENSORS - Proc.*, pages 1–4, 2015. doi: 10.1109/ICSENS.2015.7370258.

요약

도시 대기 질 측정의 시공간 해상도를 증가시키기 위해 미세먼지 센서가 광범위하게 배치되고 있다. 고해상도의 미세먼지 측정을 위한 현실적인 대안으로 저가형 미세먼지가 대표적으로 이용되고 있다. 하지만 저가형 미세먼지 센서의 측정 데이터 신뢰성에 대한 의문점은 해결되지 않고 있다. 본 연구는 저가형 미세먼지 센서의 장기간 정확도 평가를 수행하였으며, 이를 위하여 멀티 센서 플랫폼을 제작하고 이를 고신뢰도의 정부 관측소에 함께 배치하였다. 선형/비선형 추정 모델인 다중 선형회귀 모델과 인공신경망인 다층 퍼셉트론을 적용하여 데이터 기반 모델을 생성하였으며, 이를 통합한 추정 모델을 개발하였다. 이 방법들은 실외 배치 실험을 통해 평가되었으며 타 추정 모델과 타 연구와의 비교 분석을 수행하였다. 또한 관측소에 배치하여 생성된 데이터 기반 모델은 센서 네트워크를 통해 다른 노드에 전달하여 활용할 수 있으며, 이러한 접근에 대한 타당성 평가는 실험을 통해 확인하였다.

Keywords: low-cost sensor, particulate matter (PM), sensor network, accuracy, calibration, multilayer perceptron (MLP), multiple linear regression (MLR)
학번: *2018-22094*

Acknowledgements

학위 과정을 시작할 때만 하더라도 2년이라는 시간이 상당히 긴 시간이라고 생각했는데, 눈 깜짝할 사이에 지나가는 짧은 시간이었습니다. 스스로 부족을 깨달아 부족을 채워갈수록 더 큰 부족을 느끼는 시간이었지만, 배움이 주는 즐거움은 무엇과도 바꾸기 어려운 큰 기쁨이었습니다. 돌이켜 생각해 보면 10년의 직장 생활을 잠시 뒤로하고, 풀타임으로 학업에 매진할 수 있었던 지난 2년은 제 인생에서 가장 소중한 시간이었고, 학문을 바라보는 시각뿐만이 아니라 인생에 대해서도 새로운 시각을 가지는 귀중한 시간이었습니다.

이러한 시간과 시각을 허락해 주신 하나님께 감사드리며, 연구에 대한 엄밀함을 추구하며 누구보다 연구에 매진하여 솔선하시는 이동준 교수님께 깊이 감사드립니다. 연구 내용뿐만 아니라 연구에 대한 연구자의 자세에 대해서도 배울 수 있는 값진 시간이었습니다. 그리고 지난 시간 동고동락하며 알게 모르게 신경 써준 INRoL연구실 동료들과 서울시 과제를 함께하며 연구 안팎으로 많은 시간을 보냈던 지석, 성중 씨에게 감사합니다.

늦은 학위 과정이지만 항상 응원해 주신 어머니, 장인어른, 장모님, 누나, 매형님, 형, 형수님, 처형님, 형님, 처남, 처남댁에게 감사합니다. 특히 하연, 하음이를 딸처럼 생각하고 대해주시는 처형님과 형님께 감사드립니다.

마지막으로 함께 많은 시간을 보내지 못했지만, 아빠를 무한 사랑해 주는 귀한 딸 하연, 하음이에게 미안하고 고맙습니다. 그리고 학업에 집중할 수 있도록 저를 배려해주고 누구보다 헌신적으로 응원해 준 사랑하는 아내 근아에게 평생토록 고맙다고 말하고 싶습니다.

2020년 7월

이후창